410.712.81

Advance Practical Computer Concepts for Bioinformatics

**Final Project**

Student name: Yi-Fan Chou

**Title**

Annotation information display for 1000 genome project

**Abstract**

Using variants annotation tool – SnpEff to predict the effects of variants on genes, such as amino acid changes and more. Then saving the SnpEff output text file into SQLite3 database. Last, create a webpage using HTML5 and incorporate with Javascript and JQuery also add styling on the page using CSS. This will be able allow user to input and make the query on the database then have the final output on the page.

**Introduction**

Next generation sequencing is a very popular technique nowadays and it can help us to understand the genome information in many species. Typically, after sequencing we will get the raw fastq files and they will need to map to the reference genome (i.e. human - GRCh38) by sequencing alignment tool (i.e. BWA). After mapping, we will use variant calling tools (i.e. GATK-HaplotypeCaller) to identify SNPs and indels then we will get the VCF file afterward.

VCF is a text file format. It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. It is a file that contains all the differences between your sample and the reference genome. But we want to know more details about these variants, for example, Are they in a gene? In an exon region? Do they change protein coding (missense vs. nonsense mutation)? Do they cause premature stop codons? Here I would like to use SnpEff, which can help us to solve these questions.

SnpEff is a variant annotation and effect prediction tool on genetic variants. The inputs for SnpEff are predicted variants and usually in variant call format (VCF). SnpEff analyzes the input variants. It annotates the variants and calculates the effects they produce on known genes (e.g. amino acid changes). We will have the better understanding of our VCF file after SnpEff analysis.

**Method**

I used the subset of 1000 genome project VCF on chromosome 22 for our test file. Then run SnpEff to get the annotated VCF file.

**How to run SnpEff**

1. Follow the instruction to install SnpEff

http://snpeff.sourceforge.net/download.html

1. The command that I ran for SnpEff

Here, I am using human hg38 files

*java -jar snpEff.jar GRCh38.86 examples/test.chr22.vcf > test.chr22.ann.vcf*

1. Get the output file that contains annotation.

Then, I stored the annotated information (i.e. chromosome, position and annotation information contains in info filed) into SQLite3 database by using ipython notebook.

After that, I wrote a Python CGI scripts to get the user input and make the query on SQLite3 database. I used HTML5, Javascript and CSS that create a user interface that user can query the VCF file information from the databse, such as how many missense mutation on chromosome 22 in 1000 genome project VCF or how many variants are HIGH impact vs. LOW impact in terms of the effects on the variants? The users are able to select what analysis information they want to see and in which region of genome. Finally, I would like to display the query results on the page.

Here, I used two ways for user interface. 1. AJAX: can display the result at the same page without reloading it. 2. Jinja2: the output file will direct to template html file.

**Results**

AJAX method: I’ve spent the significant amount of time to debug my script, I test my search\_product.cgi that can run on the bfx server and able to get the results and print it on the screen. However, I am not able to get the results show on web page.

Jinja2 method: I tried to write alternative script that use Jinja2 for displaying the results on template page. Since the new python version that you helped me to install doesn’t contain the Jinja2 module and I am running out of time for asking you to install the Jinja2 module. I would really hope if you can still install the Jinja2 module and run my code and hopefully you will see the results display on the page.

**File contain**

AJAX method: search\_project.cgi, project.html, js/search.js, css/project.css

Jinja2 method: project\_jinja2.cgi, project\_search\_jinja2.html, project\_results\_jinja2.html, project\_jinja2.css

**Reference**

http://snpeff.sourceforge.net/

A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.", Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. Fly (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672

Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift", Cingolani, P., et. al., Frontiers in Genetics, 3, 2012.